Experimental application of probability mathematics to predictive intelligence estimates reveals a disciplinary potential.

A THEOREM FOR PREDICTION

Jack Zlotnick

Philosophy, wrote critic and educator Mortimer Adler, is the process of entertaining any idea as merely possible. This act of tentative acceptance is the good beginning in intelligence analysis. The desirable end is a correct evaluation of the several hypotheses' comparative merits.

Seldom is the evidence so determinative as to clinch the case for a single hypothesis. Usually, as it accumulates, it only changes the position of one hypothesis or another on the probability scale. Surprise attack is more likely or less likely today than it was a week ago; a Sino-Soviet break in diplomatic relations is more probable or less probable now than before; it is becoming more doubtful or less doubtful that the Labor government's position against pound devaluation can withstand the next speculative run on sterling.

Since intelligence judgments are so often probabilistic, does it follow that the mathematical theory of probability offers intelligence valid pointers on logical method? Promising research with relevance to this question, some of it government-financed, has been done by psychology faculties in university laboratories. The main aim of the psychologists has been to compare intuitive judgments about hypotheses with the results that would be given by a mathematical model based on probability theory. Borrowing from these experiments, CIA's Office of Current Intelligence in the summer of 1967 designed a mathematical simulation of predictive intelligence analysis in crisis situations of recent history.

The mathematical model derives from an equation, familiar to students of probability theory, named after Reverend Thomas Bayes, who first formulated it in the eighteenth century. The following exposition of Bayes' Theorem does not require mathematical sophistication of the reader; it assumes only that his learning blockages do not include an ingrained antipathy to any kind of numerative idea.

Bayes' Theorem

A good entry point for the discussion is the concept of probability as it is used in mathematics. In the absence of certainty, the probability that an event will occur (or has occurred, if past occurrence is the matter at issue) has a decimal or fractional value between zero and one. Thus the probability is .7 that a red poker chip will be picked in a random drawing from a box containing ten chips, seven red and three blue. A rational gambler would give no more than $7 for a raffle ticket that paid $10 upon the random drawing of a red chip from the box.

In the idiom of wagers, the term odds is often used instead of probability. The odds favoring the random selection of a red chip over the random selection of a blue one set the probability of the first event against the probability of the second. The odds of seven to three in this case are represented mathematically as the fraction obtained by dividing the .7 probability of drawing a red poker chip by the .3 probability of drawing a blue one.

New evidence affects a gambler's estimate of probabilities or odds. Suppose there are two large boxes filled with red and blue poker chips. In one the ratio of red chips to blue is 60 to 40; in the other it is 40 to 60. One of the boxes is set before a gambler, but he is not told which. He can therefore give no better than even money that its color mix is predominantly red or blue. Allow him to draw some of the chips, however, and he will then make a more confident choice between the two color-mix possibilities. The more chips he draws, the better the odds he will offer in favor of this choice.

This is precisely the setting of recent laboratory experiments at the University of Michigan and other centers. College students, serving as the test subjects, were required to give their gambler's judgments of the odds after successive drawings of poker chips, and these judgments were compared with the odds obtained by using Bayes' Theorem.

In more simplified notation than is commonly used in the textbooks, the equation of Bayes' Theorem can be written:

$$R = PL$$

R, standing for revised odds, represents the odds favoring one hypothesis over another after consideration of the latest evidence (in this case, the color of the poker chip most recently drawn). P stands for the prior odds, those prevailing before this evidence turned up. L, the weight of the evidence that changes the odds, stands for likelihood ratio (referred to sometimes in the literature as Bayes' Factor).

The likelihood ratio compares the probabilities of the occurrence of an event under alternative hypotheses. Suppose the evidence in the poker chip experiment is the selection of a red chip on the first drawing. There is a .6 probability of this happening under the hypothesis that 60 percent of the chips in the box are red. There is only a .4 probability of its happening under the hypothesis that the drawing is from the other box, where only 40 percent of the chips are red. So the likelihood ratio for the occurrence of this red drawing is .6 divided by .4, or $\frac{3}{2}$.

The prior odds, P—here $\frac{1}{1}$ for even money—are multiplied by this L to get the revised odds after the first drawing. The revised odds then become the prior odds on the second drawing, and so on. Suppose the gambler draws 12 red and 8 blue poker chips in the first 20 drawings, replacing the chip in the box after each drawing. Calculation will show that he could give better than 5 to 1 odds in favor of the hypothesis that he has been drawing from the box with the 60-40 red-blue color mix. If the first hundred drawings are 56 red and 44 blue he could give well over 100 to 1 odds in favor of this hypothesis.

Significance for Intelligence

He could and he would if he reasoned like a mathematician and had the capital to finance many wagers of this sort. Otherwise he would probably shrink from the degree of certainty implied by such high odds. The students in the University of Michigan experiments did give more confident odds the more drawings they had to go on. They did not, however, move as far from their original one to one odds as Bayes' Theorem would have justified. They did not, in other words, make the most of their inconclusive data. Like intelligence estimators in some parallel situations, they hesitated to move very far very fast from prior norms.

Similar overly conservative estimates were obtained in University of Michigan experiments simulating intelligence analysis. A set of six hypotheses was set before the test subjects—five of different imminent war situations and a sixth of peace. A scenario of events provided successive increments of evidence bearing on these hypotheses. For each increment the test subject gave five likelihood ratios expressing his opinion of how much more likely the event would be under each of the war hypotheses than under the peace hypothesis.

The test subjects of course differed among themselves in their judgment of the proper likelihood ratios. But the most noteworthy feature of the experiment was that their conclusions were not consistent with their own readings of the evidence. Like the subjects in the poker chip experiments, those working with intelligence scenarios were very conservative in their final estimates. When their likelihood ratios implied, according to Bayes' Theorem, odds of 19 to 1 in favor of a war hypothesis, their own blend of intuition and reasoning resulted typically in odds of 2 to 1. When the scenario was changed and mathematical calculations would have given 19 to 1 odds favoring peace, they came up with odds in the neighborhood of 6 to 1.

What Bayes' Theorem thus does for intelligence is to offer a mathematical test for internally consistent analysis. The rigor of mathematical logic is no indispensable aid when analysis is largely deductive, proceeding from such general propositions as "The USSR appreciates how dangerously provocative would be its shipment of strategic missiles to Cuba." The instructed intellect's naked eye, so to speak, is keen enough to follow the thread of deductive thought and to detect the more tenuous strands of the argument.

The case for mathematical assistance is stronger when analysis is more a process of inductive inference, proceeding not from a few general propositions but from many particulars. Mere verbal exposition is then less likely to ensure against fallacy and non-sequitur. Intelligence on such occasions is well advised by Francis Bacon's injunction that "the mind itself be from the very outset not left to take its own course but guided at every step; and the business be done as if by machinery." Bayes' Theorem is the kind of mechanistic aid to the intellect that Bacon here idealized.

Using this aid, the intelligence analyst does not address himself directly to the merits of hypotheses. His procedures for estimation require him to postulate, not debate, the truth of opposing hypotheses. Bayes' Theorem thus helps him get around one of his most troublesome pitfalls—his human tendency to hold fast to his prior estimate when uncommitted opinion would go along with a change. And it helps spare the estimator the labor of fighting other biases besides his own.

The Reliability Problem

In the university experiments the test subjects were in no doubt about the color of each chip they drew; nor did they have to question

the evidence set before them in the intelligence scenarios. The CIA experiment, however, incorporated a probability element to reflect the frequent uncertainties in the workaday intelligence world about the accuracy of reports from the field. The result was a modification of the Bayesian equation.

The modified equation was worked out by analogizing from the poker chip experiments. Suppose that the test subject, instead of drawing poker chips out of the box himself, turns his back and gets his information, sometimes accurate and sometimes not, from an assistant. Suppose also that he has some reasonable basis for estimating the probability of correct reporting, perhaps the assistant's past record.

Call this probability of correct reporting the reliability rating. A 30 percent reliability rating would mean that 30 percent of the reports with such a rating are true, in the rater's opinion, and the other 70 percent are false.

False reports are of two kinds. One is bereft of any corresponding fact, the utter fabrication for example. Such a report would be the assistant's announcement of a red poker chip when he had actually picked nothing at all out of the box. If the report has a probability of being false in this sense, the required modification of the equation is only to make the reliability rating (r) an exponent of the likelihood ratio:

$$R = PL^r$$

The second kind of false report is one which deliberately or innocently confuses one event with another, for example the assistant's announcement of a red chip when it was in fact blue. For reports estimated to have a probability of being false in this sense, the required modification of the equation becomes perhaps too involved to explain in a non-mathematical journal, but the mathematics is not really difficult.

The problem of the reliability rating does not enter into all appraisals of evidence. Reliability ratings are unimportant for much of the evidence received through technical collection. Nor are they necessary in intelligence appraisals of propaganda evidence, provided the analysis turns on the reasons why statements were made rather than on their truth or falsity. But the problem may well loom large in the event of garbles from technical collection and in the evaluation of reports received from human sources; and so the analyst must be at special pains to understand the very restricted meaning of the

rating. It is in no way affected by the content of a report but represents only an appraisal of source reliability, insofar as one can be made on the basis of such considerations as the amount of cloud cover in photography or the past record of clandestine human reporters. The pitfall to skirt with utmost care is the reliability rating that is nothing better than the analyst's prejudgment about the hypotheses. If his P or his R, in other words, affects his r, the analyst can find himself in a circular rut from which no mathematics can rescue him.

In real-life intelligence analysis perhaps no analyst can altogether separate his biases about the hypotheses from his appraisals of source reliability. When the credibility of some item of evidence is crucial for final conclusions, therefore, the analyst had best take a detour around the reliability issue. A case in point is the Cuban refugee report that alleges the sighting of strategic missiles near Havana. The intelligence estimator examining the hypothesis of imminent strategic missile shipments from the USSR to Cuba can hardly assign a reliability rating to this report. If he did, he would probably be putting into his analysis a judgment about credibility that is precisely the answer he wants to get out of his analysis.

To exclude altogether this refugee report and others like it from his body of evidence, however, would put the estimator into the untenable position of giving no more weight to a hundred such reports than one. His recourse is to appraise such reports much as he appraises propaganda evidence, eschewing judgment about truth or falsity. His likelihood ratio then represents only his opinion of how much more likely it is that unsubstantiated evidence of this sort would appear under the hypothesis of strategic missile shipments than under another hypothesis. This way out of the difficulty is admittedly not the most elegant of solutions, and possibilities of other methodological options are being explored.

The Cuban Missile Estimate

One test of the CIA mathematical model, a simulation of analysis just before the Cuban missile crisis, has been completed. Two intelligence exercises were simulated. One is an estimative study in mid-September 1962, when a National Intelligence Estimate on Cuba was in fact published. The other is an estimative review as of three weeks later.

The analysis sets up two mutually exclusive hypotheses. Hypothesis one is that the USSR will soon ship strategic missiles (MRBM, IRBM, or ICBM) to Cuba. Hypothesis two is that the USSR will not go so far as to ship strategic missiles, despite the sharp upsurge of military aid to Havana in the summer of 1962. The task calls for estimation of the odds favoring hypothesis one over hypothesis two.

The background of the missile crisis reaches back at least to February 1960, when a visit to Cuba by Soviet First Deputy Premier Mikoyan ended the year of Soviet reserve that followed Castro's seizure of power. In the wake of Mikoyan's visit, several economic assistance agreements were signed and Soviet deliveries of armaments commenced, giving the Cubans armored, artillery, anti-aircraft, and anti-tank capabilities appropriate for defensive and internal security purposes. The Soviets withheld the obsolescent IL-28 jet light bombers and more advanced weapons that it was supplying to other countries.

Up to 1962 more than 200 agent and refugee reports alleged the presence of missiles in Cuba. Aerial photography failed to confirm any of these reports. The Soviet Union to this point had not shipped strategic missiles to any foreign country, Communist or non-Communist.

This background information is useful only for establishing reasonable starting odds. As of January 1962, one to ten odds are postulated in favor of hypothesis one (in everyday parlance, ten to one against it). The mathematical analysis then proceeds to determine and apply likelihood ratios and reliability ratings for the evidence appearing from January 1962 on. This process, carried out in 1967 with the 1962 evidence, produces three to one odds as of mid-September 1962 against Soviet emplacement of strategic missiles in Cuba.

The mathematical calculations of 1967 thus support the estimate published in 1962. They also show, however, that the odds are shifting rapidly in favor of the strategic missile hypothesis. The fall in odds against the hypothesis accelerates, and by the end of the first week in October enough new evidence is in hand to make strategic missile emplacements an even money bet.

The Shape of Evidence

A pioneering experiment is often as interesting for the problems encountered as for the results achieved. The principal technical problem encountered in this test trial with Bayesian method was the

identification of units of evidence. In a poker chip experiment there is no doubt about the unit of evidence; it is the drawing of a poker chip of a particular color. The intelligence analyst, however, receives many reports of events. Can he make each report rather than each event his unit of evidence?

The answer is no; at least it is negative for the mathematical model used in the Cuba test. Other models may be developed, but this particular one can tell only the significance that events, not reports, have for hypotheses. To take reports as units of evidence would overweight events on which volume of reporting is high and underweight possibly more significant events on which it is low.

Several reports about the same event are therefore treated in effect as one, and volume of reporting is reflected only in the analyst's reliability ratings. These ratings represent the probability in the analyst's mind, in the light of all the reports available to him, that his evidence is accurate.

But an event, like an atom, is made up of smaller particles, and the analyst needs to have a working rule of reason to guide him in his segmentation of the evidence. The rule is to combine items of evidence so clearly associated in content that separate appraisals would virtually be double counting. Successive photography showing progress in the construction of a surface-to-air missile site can be taken as a single unit of evidence on the operational status of the site. Broadcasts on the same propaganda theme can logically be counted as one unit of evidence rather than entered broadcast by broadcast into the mathematical processing. The following two extracts from the simulated Cuba analysis illustrate the burden on the analyst to combine his evidence fairly. The italicized head names a unit of evidence; the relevant reports are then described; the unit is appraised and given a likelihood ratio, an estimate of how much more (or less) likely it is that the unit would appear if hypothesis one (strategic missiles) is true than if hypothesis two (no strategic missiles), is right.

> *Cuban-Soviet Frictions:* On 26 March, veteran Cuban Communist Anibal Escalante was ousted from party leadership. Soviet press commentary in April endorsed the removal of Escalante but also called for an end to divisions among Cuban revolutionaries. The commentary emphasized the virtues of collective leadership. The intimation of the commentary was that the USSR was disturbed by the setback suffered by its protégés in Havana.

A June report from clandestine services, originating with a usually reliable Paris source, has Castro saying privately that he wanted to stay independent of the "men of Moscow." Castro reportedly said he felt surrounded by orthodox Communists who would resort to anything to obtain control in Cuba, "even a temporary arrangement with Washington."

Fidel Castro's brother Raul, deputy premier and minister of armed forces, arrived in Moscow on 2 July. He was met at the airport by Marshal Malinovsky, the Soviet defense minister. Raul departed on 17 July without fanfare or final communique. This lack of red-carpet farewell suggested he did not get what he wanted out of the Soviets.

. . .

Simulated Mid-September 1962 Appraisal: These indications of frictions hardly put Cuba in the character of the most reliable of Soviet allies. The frictions are evaluated as unlikely, given the assumption that the Soviets are about to ship strategic missiles to Cuba. On the other hand, the frictions seem little more likely under an alternative hypothesis that assumed sharply expanded military aid of any other sort. The evidence, therefore, carries only slight diagnostic value for contradicting the hypothesis of imminent strategic missile shipments to Cuba.

Likelihood Ratio: 1 to 1.2

Reliability Rating: .8

Hints of New Cuban Capabilities: A clandestine services report in July, sourced to a fairly reliable Cuban businessman with good contacts among Castro adherents, described Cuban naval officers as pessimistic about Cuban capabilities to resist a new invasion. Cuban army officers were said to agree but to feel that the principal danger would be over by September.

In another report, a knowledgeable Cuban was quoted as saying that the US was afraid to interfere with Soviet-flag vessels but "in September the Americans will also respect the Cuban flag."

At one point, the Cuban (Che Guevara according to one account) referred to the NATO nations as a belt of bases surrounding the Soviet Union. He was reportedly "livid" as he added that "in September Cuba is going to be the buckle in this belt."

. . .

Simulated Mid-September 1962 Appraisal: The allusion to NATO bases suggests a development consistent with the assumption of strategic missile installations in Cuba. The allusion could also have been expressed, although less probably, given the assumption of expanded military aid to Cuba that stopped short of strategic missile emplacements.

The accuracy of the reports is open to question.

Likelihood Ratio: 1.5 to 1

Reliability Rating: .5

As these two extracts indicate, the telescoping of reports sharply reduces the number of units of evidence available for mathematical processing. The reduction gravely complicates the analyst's task. The reason is that Bayesian analysis takes off from starting odds which

may be more intuitive than grounded in evidence. If many units of evidence are available, these should in time outweigh the influence of the starting odds. The rub comes when there are not many units of evidence. The prospect is then that starting odds rather than evidence will constitute the predominating influence on the final odds.

The Cuba test suggests that this problem will bedevil intelligence more often than not. Intelligence collection during the Cuban military buildup was massive, but the evidence touched on comparatively few subjects. The opportunities to increase or reduce the starting odds of ten to one against the strategic missile hypothesis did not, therefore, come thick and fast, and an analyst would want to offer his choice of hypothesis with considerable reserve. Perhaps the best he could do would be to say how much the evidence had shifted the odds since the starting date of his analysis. While this interpretation might not justify confident predictions, it could alert policymakers to the implications of recent developments.

Critique

Working with the Bayesian model, intelligence is not a blend of deduction, insight, and inference from the body of evidence as a whole. It is a sequence of explicit judgments on discrete units of evidence. Bayesian analysis can carry conviction only if the evidence itself persuades. The analysis cannot apply the additional dialectic leverage of well-reasoned generalization cast in finely finished phrase.

This necessity to work with a hard base of evidence limits the prospective usefulness of Bayesian method. Current evidence in many situations carries little weight for longer-term estimation. Even for short-term prediction, the base of available evidence may be too small a foundation to support by itself the estimative structure that intelligence must often put together for the high councils of government. A forecast of foreign reaction to a postulated course of U.S. action probably has some evidence to go on but not much, at least not until the United States gets nearer the decision to take the postulated action.

Would it be worth while, then, for estimates of the future to include such an interpretive tabulation of all units of evidence as in the Cuban simulation? There is much to be said for requiring such a tabulation in all cases. Bayesian method is helpful not only for its rules to assure valid induction but also for its duress on the analyst to separate fact from opinion. Even if the analyst does not follow through with mathematical processing, his analysis should be the better for his labor in

poring over details of evidence and for the resulting higher level of explicitness in his working materials. Should the tabulation of relevant evidence be embarrassingly short, both analyst and reader are alerted to the weakness of the evidential base and to the pivotal position of a priori judgment in the estimate.

To argue for evidence, however, is to knock on an open door. Everyone would like to appeal to the verdict of evidence. The deep skepticisms are not about the virtues of evidence but about the practicality of representing evidence with mathematical precision. It is one thing to work with probabilities of drawing a red poker chip from a box with a given color mix of chips. Is it not quite another thing to work with likelihood ratios and reliability ratings that are personal opinions about the probabilities? The underlying data in the one case are numerical counts, and all the experts are agreed on the rules for assigning probability values to such data. In the other case, the probabilities are subjective judgments and tentative besides. If the intelligence analyst says that an event is twice as likely to happen if one hypothesis is true than if another hypothesis is true, does he really want that figure to be taken literally? And if he says the chances are only four out of ten that a source is reporting accurately, does he want precisely this opinion about the source and no other to count in the basis of his final conclusions?

The question is almost its own answer. The likelihood ratios and reliability ratings do no more than suggest roughly how the analyst is weighing evidence in his own mind. Mathematical processing in real-life intelligence analysis ought not, therefore, to restrict itself to one set of likelihood ratios and reliability ratings. It should rather involve several passes over the evidence with different sets of figures.

The processing would thus show the sensitivity of final conclusions to variation in appraisals of the evidence. Suppose one or two mixes of likelihood ratios and reliability ratings led to a conclusion that contradicted those given by the other passes over the evidence or that contradicted the intelligence consensus reached by conventional analysis. It should then be incumbent on the analysts to determine the reason for this contradictory conclusion. They might decide in the end to rule against it on the ground that it was based on unreasonable weighting of the evidence. But if they felt the weighting was not beyond the bounds of reason, they might decide to rethink the whole subject.

Mathematical processing will not become an alternative to present methods of intelligence analysis. It will become a reliability check on present methods. It will help show the plausibility of conclusions which the intelligence analyst would not otherwise recognize as compatible with the evidence and his own inner logic. It will tell the analyst: if you interpret the evidence in this way, then here is the conclusion you should probably reach. Often the mathematics will be persuasive.